SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
COMPUTATIONAL MODELS
BAYESIAN INFERENCE
DATA INTEGRATION
PROBABILISTIC DATA NETWORKS

Received
21 September 2012

Accepted
7 December 2012

Published
21 January 2013

Correspondence and requests for materials should be addressed to L.C. (lnchen@sibs.ac.cn); X.S.Z. (zxs@amt.ac.cn) or H.Z. (hrzhou@sibs.ac.cn)

* These authors contributed equally to this work.

APG: an Active Protein-Gene Network Model to Quantify Regulatory Signals in Complex Biological Systems

Jiguang Wang[1,2,4,6]*, Yidan Sun[1,3]*, Si Zheng[2], Xiang-Sun Zhang[4], Huarong Zhou[1] & Luonan Chen[1,4,5]

[1]Key Laboratory of Systems Biology, SIBS-Novo Nordisk Translational Research Centre for PreDiabetes, Shanghai Institutes for Biological Sciences, CAS, Shanghai 200233, China, [2]Beijing Institute of Genomics, CAS, Beijing 100029, China, [3]Key Laboratory of Human Functional Genomics of Jiangsu Province, Nanjing Medical University, Nanjing 210029, China, [4]National Center for Mathematics and Interdisciplinary Sciences, CAS, Beijing 100190, China, [5]Institute of Industrial Science, University of Tokyo, Tokyo 153-8505, Japan, [6]Department of Biomedical Informatics and Center for Computational Biology and Bioinformatics, Columbia University, New York, NY 10032, USA.

Synergistic interactions among transcription factors (TFs) and their cofactors collectively determine gene expression in complex biological systems. In this work, we develop a novel graphical model, called Active Protein-Gene (APG) network model, to quantify regulatory signals of transcription in complex biomolecular networks through integrating both TF upstream-regulation and downstream-regulation high-throughput data. Firstly, we theoretically and computationally demonstrate the effectiveness of APG by comparing with the traditional strategy based only on TF downstream-regulation information. We then apply this model to study spontaneous type 2 diabetic Goto-Kakizaki (GK) and Wistar control rats. Our biological experiments validate the theoretical results. In particular, SP1 is found to be a hidden TF with changed regulatory activity, and the loss of SP1 activity contributes to the increased glucose production during diabetes development. APG model provides theoretical basis to quantitatively elucidate transcriptional regulation by modelling TF combinatorial interactions and exploiting multilevel high-throughput information.

High-throughput technologies, such as DNA microarray, deep sequencing, yeast 2-hybrid, and protein mass spectrometry, generate tremendous amount of data at genome-wide scale and also at different molecular levels[1–5], which provides snap shots of the cells under different conditions. To explore rich information of such high-dimensional data, computational methods are needed for identifying key genes, such as transcriptional factors (TFs), and also for inferring their upstream-regulation and/or downstream-regulation interactions. Differential expression analyses are widely used to find hot-spot genes or proteins. However, the results derived simply based on only the abundance of mRNAs or proteins sometimes show low accuracy or even lead to wrong conclusions[6]. For example, a TF which regulates its target genes by binding to DNA with its cofactors may change its function or activity by interacting with different cofactors or rewiring its network even without any alteration of its mRNA or protein expression level. Thus, although many of TFs play central roles during a perturbed biological process, they exhibit no significant changes at mRNA or protein levels and thereby are frequently overlooked by scientists. On the other hand, molecular interactions or regulatory relations such as TF-TF interactions or TF-target gene regulations found by *in vitro* experiment in one condition do not always exist in other conditions. As a result, an imperative and challenging task remains to quantify TF activities and reveal their interactions so as to elucidate the key regulatory processes behind physiology and pathology[7–9], by making better use of the multilevel high-throughput data.

TFs are usually key regulators of cell fate or biological processes. In recent years, several research works have studied TF functionality, i.e., TF activities, through mRNA expression profiling. Liao et al. developed a statistical assumption-free approach, named Network Component Analysis (NCA), to infer TF activity, which reflects the ability of TFs to regulate the transcription of mRNAs[10]. Meanwhile, Carro et al. inferred *de novo* TF-target interactions by an information theoretical approach, named ARACNe, and then discovered the master regulators of mesenchymal transformation by computing the statistical significance of the overlap between the targets of each TF and the MGES genes by Fisher's exact test[11]. Both NCA and ARACNe identify TF activities by using target gene expression as their reporter, i.e., TF downstream-regulation information. Those types of methods,

providing a computational way to discover key regulators even without abundance changes of their mRNA levels, dramatically improve our understanding of underlying functions for those "hidden" or "unobservable" key TFs. Nevertheless, since many TFs, regulating their target genes, change their functional roles by interacting with different cofactors, TF activities are mainly determined by protein interactions among TFs and their cofactors, i.e., TF upstream-regulation information, rather than the downstream-regulation information. Therefore, to infer TF activity in an accurate manner, it is important to exploit TF upstream-regulation information (e.g., expression levels of TF cofactors) in addition to TF downstream-regulation information (e.g., expression levels of TF target genes).

Enlightened by this fact, we extend the concept of TF activity described by Liao et al.[10] as an integrative index reflecting not only cooperativity of transcriptional factors with cofactors but also the ability of transcriptional complexes to regulate the transcription of mRNAs. That is, we propose a novel method based on a causal "cofactor-TF-target" cascade, called Active Protein-Gene (APG) network model, by integrating both upstream-regulation and downstream-regulation structures of TFs to quantitatively infer not only regulatory strengths of TFs but also their regulatory network structure. Unlike the previous approaches mainly using the mRNA information of TF targets (i.e., TF downstream-regulation information), APG integrates both TF upstream-regulation and downstream-regulation information, thereby requiring less samples and TF-target connectivity information to ensure the accurate inference of TF activities and network structure. Specifically, we first theoretically prove that there is a unique solution for APG model even without prior knowledge based on a graphical model and matrix factorization theory, which significantly extends previous methods based only on downstream-regulation information. We also numerically show that APG always has higher accuracy than the downstream-regulation information methods, for the cases not only with less prior knowledge but also with higher white noise. Second, we examine the performance of APG model by applying it to liver microarray data from type 2 diabetic GK rats and Wistar controls[12]. GK colony is established by more than 30-generation repeated breeding of Wistar rats with blood glucose in upper limit of normal distribution for glucose tolerance[13]. Spontaneous hyperglycemia in GK colony is not established by single gene mutation, thus GK rat is considered as one of the best animal models for studying type 2 diabetes, which represents majority of all cases in diabetes[13]. Being a valuable tool offering sufficient commonalities to study human Type 2 diabetes, finding key hidden TFs with changed activities and also their regulatory network in diabetic GK rats will provide pathological hypotheses for human disease. Actually, by applying APG to animal models, we identify several hidden TFs, especially SP1, during development of diabetes, which cannot be detected by the traditional differential expression scheme due to no significant differential expression for SP1. Third, APG discovers the rewiring network of TFs and their regulatory targets by comparing GK rats and Wistar controls. In particular, for the first time, strong correlations of E2F1 and BAHD1 to GCK in Wistar rats are revealed, in contrast to weak correlations of SP1, E2F1 and BAHD1 to GCK in GK rats. Finally, we conduct the biological validation experiments, in which we confirm not only the relevant network structures of the TFs but also SP1 as a key TF with the hidden regulator activity. In particular, we show that the loss of SP1 activity contributes to the increased glucose production during diabetes development in GK rats.

Results

Active Protein-Gene (APG) network model. In contrast to models mainly using known or *de novo* predicted TF downstream-regulation information, the APG model integrates both TF upstream-regulation and downstream-regulation information by exploiting causal cofactor-TF-target relations, i.e. protein interactions among TF and cofactors, concentrations of TF and cofactors, TF-target connectivity, expression levels of target genes, and other factors (Figure 1). For this purpose, our model is composed of three layers. The first layer consists of proteins including transcriptional factors, cofactors, or other proteins related to the activity of TFs such as kinase and phosphatase, which is the layer of the cause of TF activity. The second layer represents the TF activity, which shows cooperativity of transcriptional factors with cofactors or the ability of TFs to activate (or repress) target gene transcription. This is a hidden layer acting as a modulator to transform signals of TFs and their cofactors to their target genes, i.e. values in this layer cannot be directly experimentally measured. The third layer including expression levels of different target genes is the effect or result of TF activity.

We represent concentrations of proteins affecting TF, TF activities, and expression level of targets with random variables $Q=(Q_1,\cdots,Q_L,\cdots,Q_K)$, $P=(P_1,\cdots,P_L)$, and $E=(E_1,\cdots,E_N)$, respectively. Here, $Q_1,\cdots,Q_L$ are concentrations of TF, and $Q_{L+1},\cdots,Q_K$ are concentrations of proteins modifying TF. According to the dependent and conditional independent feature of Bayesian network, we can get the joint distribution

$$\Pr(Q,P,E)=\prod_{k=1}^{K}\Pr(Q_k)\prod_{l=1}^{L}\Pr[P_l|pa(P_l)]\prod_{n=1}^{N}\Pr[E_n|pa(E_n)] \quad (1)$$

where $pa(X)$ stand for all parents of node X in Bayesian network. Clearly, $pa(P_l)$ or $pa(E_n)$ are a subset of Q or P.

In this model, random variables can be both discrete and continuous. Here we consider continuous case. A natural choice for representing continuous variables is the use of Gaussian distribution. According to the linear Gaussian model, we have the conditional density of X given its parents

$$\Pr(X|A_1,\cdots,A_k)\propto N\left(\sum_i \lambda_i a_i,\sigma^2\right)$$

where $A_1,\cdots,A_k$ are the parents of X; $N(\mu,\sigma^2)$ is the density function of the normal distribution with mean μ and standard variation σ, and $a_1,\cdots,a_k$ are the observations of $A_1,\cdots,A_k$ respectively. λ_i is the effect strength of the ith variable. Here, the variation σ is independent of the value of parents. Since the activities of TFs depend on concentrations of both TFs and corresponding cofactor proteins, we have

$$\Pr[P_l|pa(P_l)]\propto N\left(\sum_k \beta_{lk}q_k,\sigma^2_{P_l|pa}\right) \quad (2)$$

where $l=1,2,\cdots,L$; q_k is the observed value of Q_k; β_{lk} is the effect strength of the kth protein on the lth TF. Similarly, we have

$$\Pr[E_n|pa(E_n)]\propto N\left(\sum_l \alpha_{nl}p_l,\sigma^2_{E_n|pa}\right) \quad (3)$$

where $n=1,2,\cdots,N$; p_l is the potential observed value of P_l; α_{nl} is the regulatory strength of the lth TF on the nth target gene.

Then according to formulas (1)–(3), we have

$$\Pr[Q,P,E]\propto\prod_{k=1}^{K}N\left(\mu_{Q_k},\sigma^2_{Q_k}\right)\prod_{l=1}^{L}N\left(\sum_k\beta_{lk}q_k,\sigma^2_{P_l|pa}\right)\prod_{n=1}^{N}N\left(\sum_l\alpha_{nl}p_l,\sigma^2_{E_n|pa}\right) \quad (4)$$

where α, β are parameters on network structure; μ, σ are parameters on data, and p, q are respectively absent and observed data.

To infer the optimal parameters or network structure of the graphical model, we use Maximum Likelihood Estimation (MLE) to maximize the probability of observed data[14]. Particularly, if

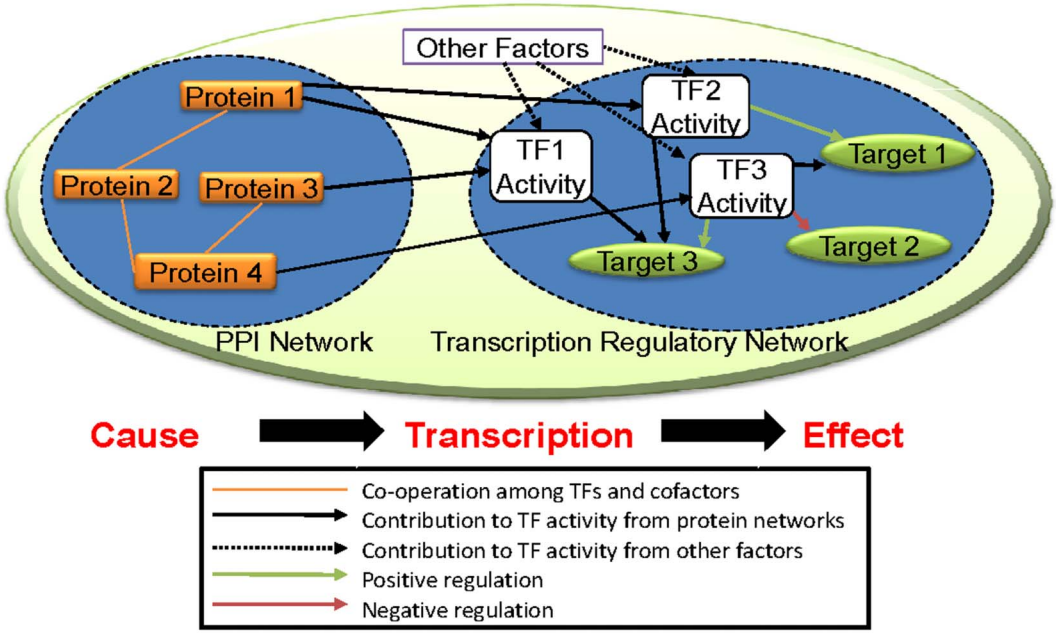

Figure 1 | Schematic diagram of APG model. There are causal relations between protein interaction and transcriptional regulation. APG model integrates both TF upstream-regulation and downstream-regulation information by exploiting causal "cofactor-TF-target" relations, i.e., protein interactions among TF and cofactors, concentrations of TF and cofactors, TF-target connectivity, expression levels of target genes, and other factors.

there are M independent observations of Q and E, we let $\Theta=\left(\alpha,\beta,\mu_{Q_k},\sigma_{Q_k},\sigma_{E|pa},\sigma_{P|pa}\right)$, then the MLE is to maximize logarithm likelihood function $L(\Theta|Q,E)=\log\prod_{i=1}^{M}[\Pr(Q_i,P,E_i)]$, where Q_i,E_i are respectively observed values of Q,E in ith sample. To do that, the hard-assignment version EM strategy[14] is carried out as shown in Supplementary Text S1. Thus, we can obtain the network structure (α,β) and TF activities (P) provided that we get Θ.

Conditions of unique solution for TF activities and network structure. One important question to apply AGP model is whether or not the solution for TF activities and network structure is unique. The mathematical framework of APG in a matrix form can be summarized by

$$E=AP \tag{5}$$

$$P=BQ \tag{6}$$

where (5) considering TF-downstream-regulation information of P is actually the framework of NCA. Specifically, E being measurable is a $N\times M$ matrix representing expression level of N genes under M different conditions. A is TF downstream-regulation structure which is the $N\times L$ connectivity matrix representing the regulations between L TFs and N targets. P being immeasurable is an $L\times M$ matrix representing TF activity at these M conditions. On the other hand, (6) considers TF-upstream-regulation information of P. Specifically, B is TF upstream-regulation structure which represents the effect of protein on TF activity including modifications of proteins and TF concentrations; Q are the concentrations of TF and activities of modifying proteins. Clearly, by substituting (6) into (5), we have $E=ABQ$. If E or Q as well as partial elements of A and B are given, the variables to be determined are the remaining unknown elements of A and B.

Then, if we assume that ith TF does not directly affect the activity of jth TF for $i\neq j$, we can prove that the solution of (5–6) is unique up to a scaling factor when A has a full-column rank and Q has a full-row rank (refer to Corollary 2 in Supplementary Text S2). That is to say by integrating both TF upstream-regulation and downstream-regulation structures, APG ensures the unique solution for TF activities and their network structure even with no requirement on prior knowledge of TF-target information or kinase-TF information, which extends the results of NCA in (reference 10).

More generally, if the assumption that ith TF directly affects the activity of jth TF for some $i\neq j$ is not invalid, we derive another series of conditions to ensure the unique solution for APG, which also require less TF-target information than the traditional (downstream-regulation) methods (Supplementary Text S2 Theorem 1 and Corollary 1).

Numerical experiments. To illustrate the effectiveness of our method, we firstly constructed two simple networks. The first one contains one protein, one TF and two targets with prior knowledge satisfying NCA criterions (Supplementary Figure S1), and the second one contains two TFs, three candidate cofactors, and three candidate targets with prior knowledge not satisfying NCA criterion ii (Supplementary Figure S3). However, both of the two examples satisfy the conditions of APG. We applied APG to infer TF activity and network structures from 10 samples randomly generated by normal distribution. All TF activities and network parameters were successfully deduced by APG even without any prior information (Supplementary Figures S2 and S4). To quantitatively compare the efficiency of APG and the downstream-regulation method on these two examples, we defined the Root Mean Square Deviation (RMSD) between the inferred and real activity for TF_j as follows

$$RMSD_j=\sqrt{\frac{1}{m}\sum_k\left(y_k^{TFA_j}-\hat{y}_k^{TFA_j}\right)^2},$$

where m is the number of samples, y^{TFA_j} is the real activity of TF_j and $\hat{y}^{TFA_j}$ is the inferred activity. Then we calculate RMSD for APG and the downstream-regulation method at different noise level (σ). Supplementary Figure S5 A and B show the results of APG and the downstream-regulation method on the first and second networks, respectively. Both the downstream-regulation method and APG work well on the first network when σ is small (RMSD is small). With the increase of noise, accuracy (mean value of RMSD)

and convergent stability (standard deviation) of both methods gradually deteriorate, but APG with smaller standard deviation and RMSD always outperformed the downstream-regulation method. In the second network, the downstream-regulation method resulted in big errors while APG still gave the solution with reasonable accuracy and convergence. Then we compared convergent rate for the downstream-regulation method and APG within given number of iteration steps. As shown in Supplementary Figure S5 C and D, clearly APG is superior to the downstream-regulation method, showing that APG has a better convergent rate. Besides, we compared RMSD and convergent rate of APG and the downstream-regulation method for different sample sizes. Numerical experiments on these two networks also show superiority of APG comparing with the downstream-regulation method (Supplementary Figure S6).

To further illustrate the effectiveness, we constructed a more complex artificial example with 50 TFs, 50 cofactors and 200 target genes, following the workflow shown in Supplementary Figure S7. Firstly, we simulated protein concentrations of TFs and cofactors by random numeric values following normal distribution with fixed mean values and standard deviations for each protein. Secondly, the TF activities were deduced by protein concentrations and active TF-cofactor cooperation network. Specifically, to construct the active TF-cofactor cooperation network, we firstly assumed that there is a potential TF-cofactor protein interaction network for the organism we are interested in, and then removed θ% of all edges by random to generate TF activity following formula (2) with standard deviation σ. Here, the potential TF-cofactor interaction network was randomly constructed with cofactor out-degree following power-law distribution. Thirdly, we constructed the potential TF-target regulatory network, randomly produced the active TF-target network, and generated the simulated gene expression data by formula (3) with similar strategy. In this example, we tried to infer TF activity additional with active network structure, under the assumption that protein concentrations, gene expression level, potential TF-cofactor interaction network, and potential TF-target regulatory network are available data. To do that, we applied APG from 100 samples randomly generated by strategy described above. TF activities and network parameters were successfully deduced by APG in most cases. To compare the efficiency of APG and the downstream-regulation methods (i.e., downstream-regulation information based methods) on this example, we calculated RMSD for APG and the downstream-regulation methods at different inactive rate (θ) and different noise level (σ). Figure 2A and 2B show the results of APG and the downstream-regulation methods, respectively. Both the downstream-regulation methods and APG work well on the first network when σ and θ are small. With the increase of noise or inactive rate, accuracy of both methods gradually deteriorated, but APG always outperformed the downstream-regulation methods due to the integration of more information (Figure 2A and 2B).

APG model infers TF activity in diabetic GK rat model. As a proof-of-concept application, we used APG model on the diabetic GK rat data collected above. Specifically, there are total 50 samples across five time points, with five diabetic rats and five normal rats at each time point as shown in Supplementary Figure S8. For each gene or protein (here gene expression level was used to approach protein concentration), we used x_{it} to represent its expression level in the liver of GK rat i at time point t, and y_{jt} Wistar rat j at time point t. In order to figure out which TF is the master or key regulator in regulating differentially expressed genes in diabetic rat, we prepared the input data matrix for APG by calculating $e_{it} = \frac{x_{it}}{y_t}$ $= \log_2 \frac{x_{it}}{mean(y_{jt})}$. Then we got the differentially expressed levels of all TFs, cofactors, and targets genes of 25 samples across all five time points (five for each). Then we ran APG with input of 113 TFs, 121 cofactors, 335 targets in 25 samples, and obtained the TF activities of all TFs at each sample. Noticing that there are five samples at each time point, we further calculated p-value to evaluate whether TF activity is larger than zero (gain of activity in diabetic rat comparing with normal) or less than zero (loss of activity in diabetic rat comparing with normal) at each time point for each TF with T-test. Finally, we showed that fold changes of activities of 25 TFs are significantly different from the corresponding fold changes of the mRNA levels between GK and Wistar rats at more than one time points in Figure 3A. Interestingly, most of the TFs showing above with high or low activities have moderate or even opposite mRNA changes compared GK with Wistar. This indicates that, in terms of TFs, the mRNA and activity separation is not a rare case.

In particular, we found the remarkable differences between the activities and their mRNA expression levels of 3TFs, i.e., JUN, SP1 and STAT5B. The three TFs showed no significant fold changes of the mRNA expression levels but had highly significant differences in their activity levels at least 4 time points. SP1 is a housekeeping gene binding with high affinity to GC-rich motifs and regulating a variety of functions such as cell growth and apoptosis[17,18]. JUN is highly similar to the viral protein avian sarcoma virus 17, which is well known to be involved in both translocations and deletions in human malignancies[19]. A weaker DNA binding of STAT5B has been linked to defective pathways contributing to diabetes in mice[20,21]. It is well

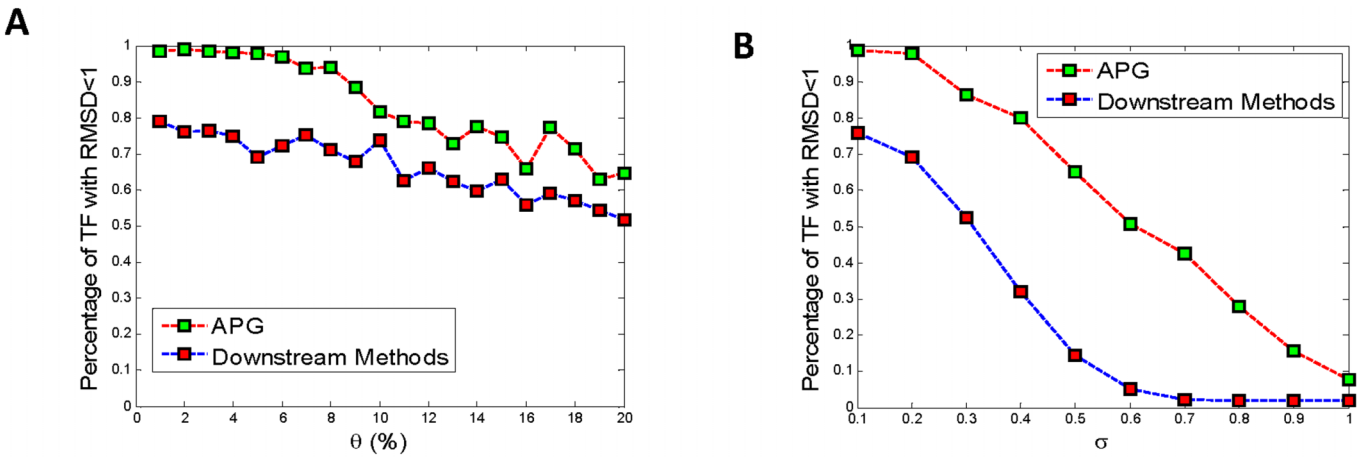

Figure 2 | Numerical experiments showing the ability of APG for inferring TF activity and reconstructing network structure. (A) RMSD is computed to measure the accuracy between predicted activity and actual activity. With different inactive ratio θ, APG always obtain better accuracy. (B) RMSD of APG and the downstream-regulation based methods with different noise levels.

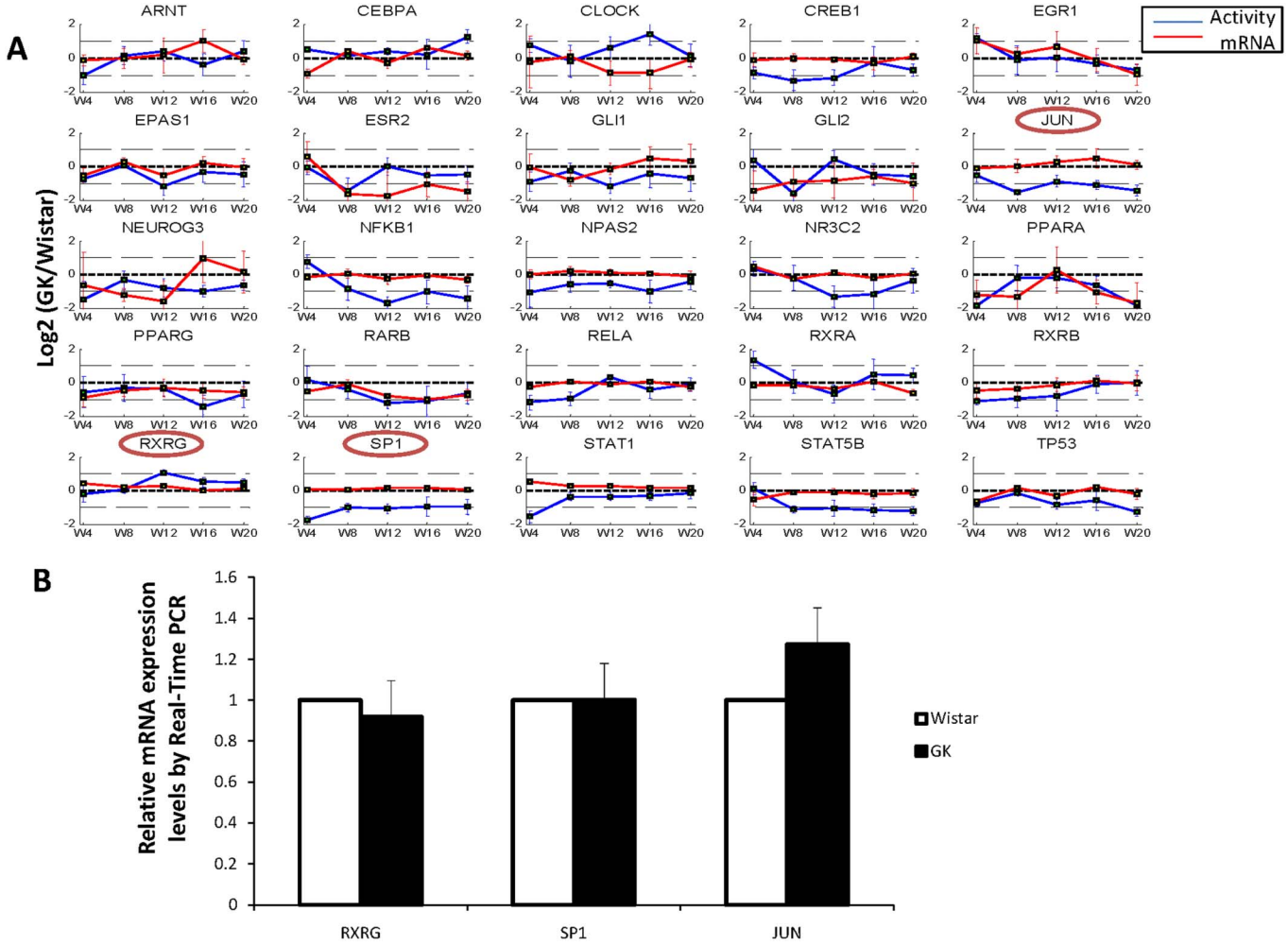

Figure 3 | Results of APG on biological data. (A) TFs with significant activity changes. We first collected the liver gene expression data from GK and Wistar rats in Gene Expression Omnibus (GEO) database (GSE 13271). For each time point, mean and standard variant were calculated on five biological repeated samples for both activity and mRNA levels of corresponding TF. W means weeks of age. (B) Relative mRNA levels of RXRG, SP1, and JUN were measured in the livers from 4-week old Wistar (n=5) and GK (n=6) rats. There were no significant changes of these three gene expressions between GK and Wistar rats.

known that diabetes is reversible during early stage, but the information during early development of diabetes is rare. Thus we picked 4 weeks of age to study hidden TFs which might contribute to development of diabetes. We first selected SP1 as the candidates of hidden regulators to further experimental verification. Since both JUN and STAT5B did not show significant TF activity changes at 4 week, we selected JUN as a TF without significant change of activity. In addition, RXRG, which only has significant differences of the activity at only 12 weeks of age, was also selected as a control. RXRG interacts with the retinoic acid, thyroid hormone, and vitamin D receptors increasing their functions[22]. As a result, we selected the three transcription factors, JUN, SP1 and RXRG, to validate them as hidden regulators and controls in 4 week diabetic GK rats.

We first checked the mRNA expression levels of the three TFs in the livers from 4 weeks old GK and Wistar rats. The real-time PCR results showed that mRNA levels of all three genes were not significantly different between GK and Wistar rats at 4 weeks, comparable to those in the microarray data (Figure 3B). In order to confirm activity levels, direct and indirect methods were used. Some TF activities are correlated to domain phosphorylation, thus their activities can be directly measured by their protein phosphorylation levels. JUN, also called c-Jun, is activated through double phosphorylation by the c-Jun N-terminal kinases (JNKs) on Ser-63 and Ser-73 within its transcriptional activation domain. Thus we measured JUN-Ser-63 protein levels in the liver of GK and Wistar rats at 4 weeks of age. As seen in Figure 4A, the phosphorylation levels of JUN-Ser-63 showed no significant difference between Wistar and GK rats, which indicated no alteration of the transcriptional activity of JUN, in consistence with APG prediction.

Since activities of many TFs cannot be tested directly, we designed an indirect way to speculate their activities. One of the most important characteristics of type 2 diabetes is the phenomenon of the increased glucose production in liver. The activity of SP1, but not RXRG, is significantly different as predicted by APG. We believe that if the predictions are true, we should certainly observe the different changes of glucose production in primary hepatocytes from Wistar and GK rats after knocking down TF by specific small interference RNA (siRNA). To test this hypothesis, we transfected siRNAs, which are specific to SP1 and RXRG, into primary hepatocytes of Wistar and GK rats, and analyzed their mRNA expression levels by real-time PCR. When SP1 mRNA levels were dramatically decreased by siRNA, glucose productions from Wistar rat hepatocytes were measured. While knocking-down SP1 mRNA levels caused a significant 14% up regulation of glucose production from Wistar rat hepatocytes, no significant changes were observed in GK hepatocytes after similar SP1 mRNA knocking down (Figure 4B and 4C). In contrast, about 90% RXRG mRNA knocking down in GK and Wistar hepatocytes did not significantly change glucose production in both GK

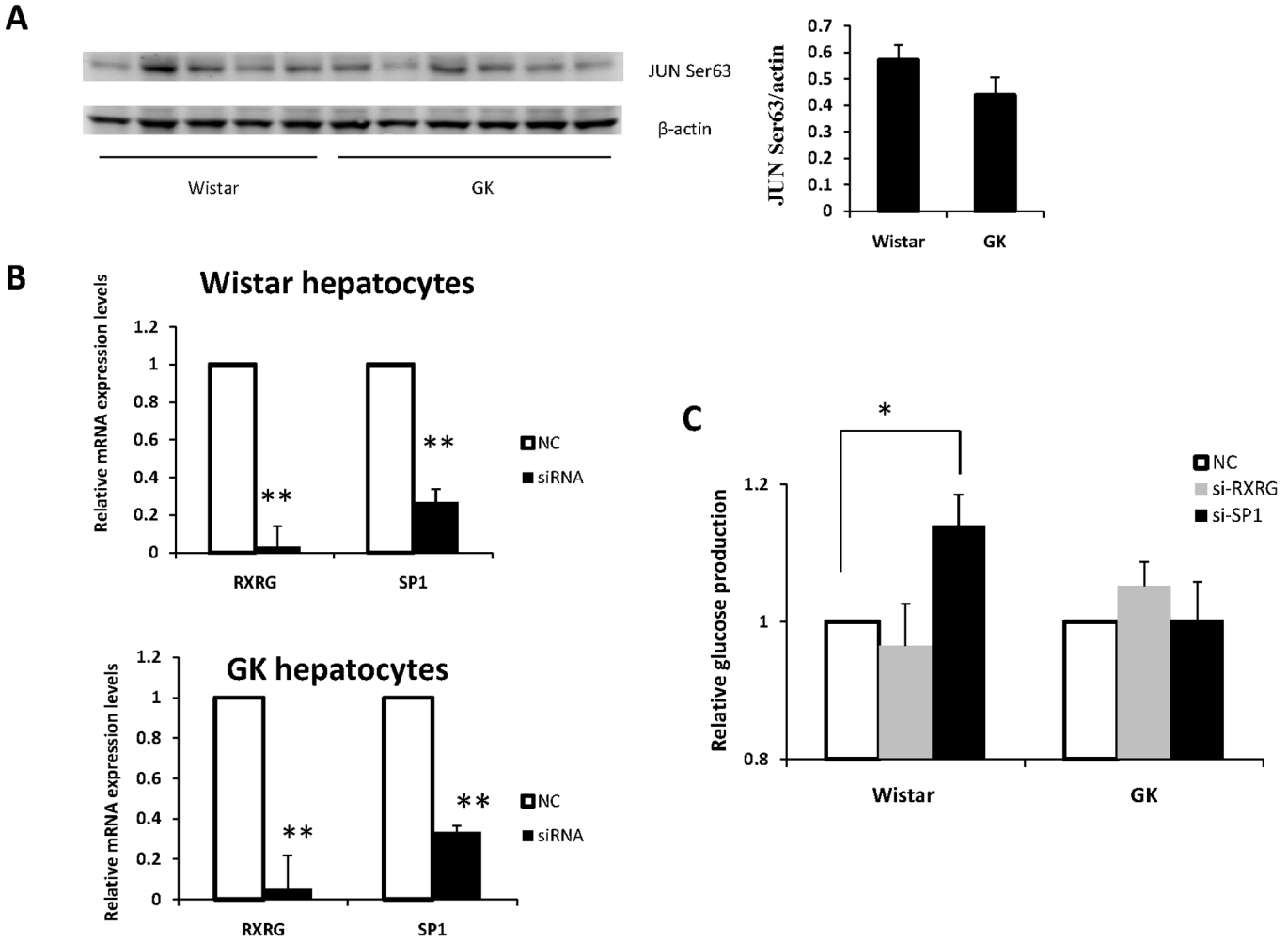

Figure 4 | Validation of activities. (A) No significant changes of JUN-Ser-63 in the liver tissues from Wistar (n=5) and GK rats (n=6). The phosphorylation levels of JUN were determined by Western Blot, and then analysed by densitometry. (B) Gene expression after siRNA knocking down. Primary hepatocytes isolated from Wistar and GK rats were transfected by negative control siRNA (NC), Rxrg and SP1 specific siRNA. The mRNA levels of SP1 and Rxrg were significantly decreased after specific siRNA transfection. The gene knocking down degrees in Wistar and GK primary hepatocytes were comparable. (C) Glucose production of primary hepatocytes isolated from Wistar or GK rats after siRNA knocking down. Knocking down SP1 mRNA levels caused a significant up regulation of glucose production from Wistar rat hepatocytes. No significant changes in glucose production were observed in GK rat hepatocytes after similar SP1 mRNA knocking down. Knocking down RXRG mRNA levels did not significantly change glucose production in hepatocytes from both GK and Wistar rats at 4 weeks of age. For Figure B and C, Error bar indicates standard deviation for each group (n=3) in three separate experiments.

and Wistar hepatocytes at 4 weeks of age, which suggests that RXRG activity is not closely related to liver glucose metabolism changes during this age. These results suggest that, in terms of glucose production, SP1 is functional in 4-week Wistar rats. Despite similar mRNA levels, the activities of SP1 are lost in diabetic GK rats at 4 weeks of age. In such ways, the activities of SP1 as a hidden regulator contributing to development of diabetes were verified by glucose production.

Network rewiring inferred by APG in development of diabetes. Identifying network rewiring in diabetes at early stage can provide valuable information on the biological mechanisms of the initiation and development for the disease. Blood glucose levels of GK rats at 4 weeks of age do not reach the established glucose criteria for the diagnosis of diabetes, thus animals at this age are considered prediabetes[23]. To infer network rewiring in 4-week old prediabetic GK rats, we firstly selected all significant TFs at 4 weeks of age, and then used them as seeds to reconstruct upstream-regulation and downstream-regulation cascades by network structure predicted by APG. Specifically, for given TF l, if not only $|\alpha_{nl}| > 0.1$ but also gene n is significantly differentially expressed between diabetic and normal rats (p-value<0.05), gene n is selected as one of the down stream-regulation targets. Similarly, if $|\beta_{lk}| > 0.1$ and cofactor k is significantly differentially expressed between GK and Wistar, cofactor k is selected as one of the upstream-regulation cofactors at 4-week old. As shown in Figure 5A, diamonds and squares represent TF activities (diamonds represent high activity while squares low activity). EGR1 and RXRA have higher activities, while TFs, such as SP1, PPARA, RXRB, ARNT, STAT1, NPAS2, RELA, and NEUROG3, have lower activities. We also constructed the causal and regulated layers of the genes presenting in circles, based on the connectivity matrix and gene expression profile as described above. TF upstream-regulations are presented by dash lines, while the solid lines indicate TF downstream-regulation signals. The mRNA levels of all genes are showed by spectrum colors from green to red (green mean lower expression level in diabetic rats, while red means higher). We firstly check the enriched GO terms by NOA (Network Ontology Analysis), which is a novel Gene Ontology tool aiming to analyze

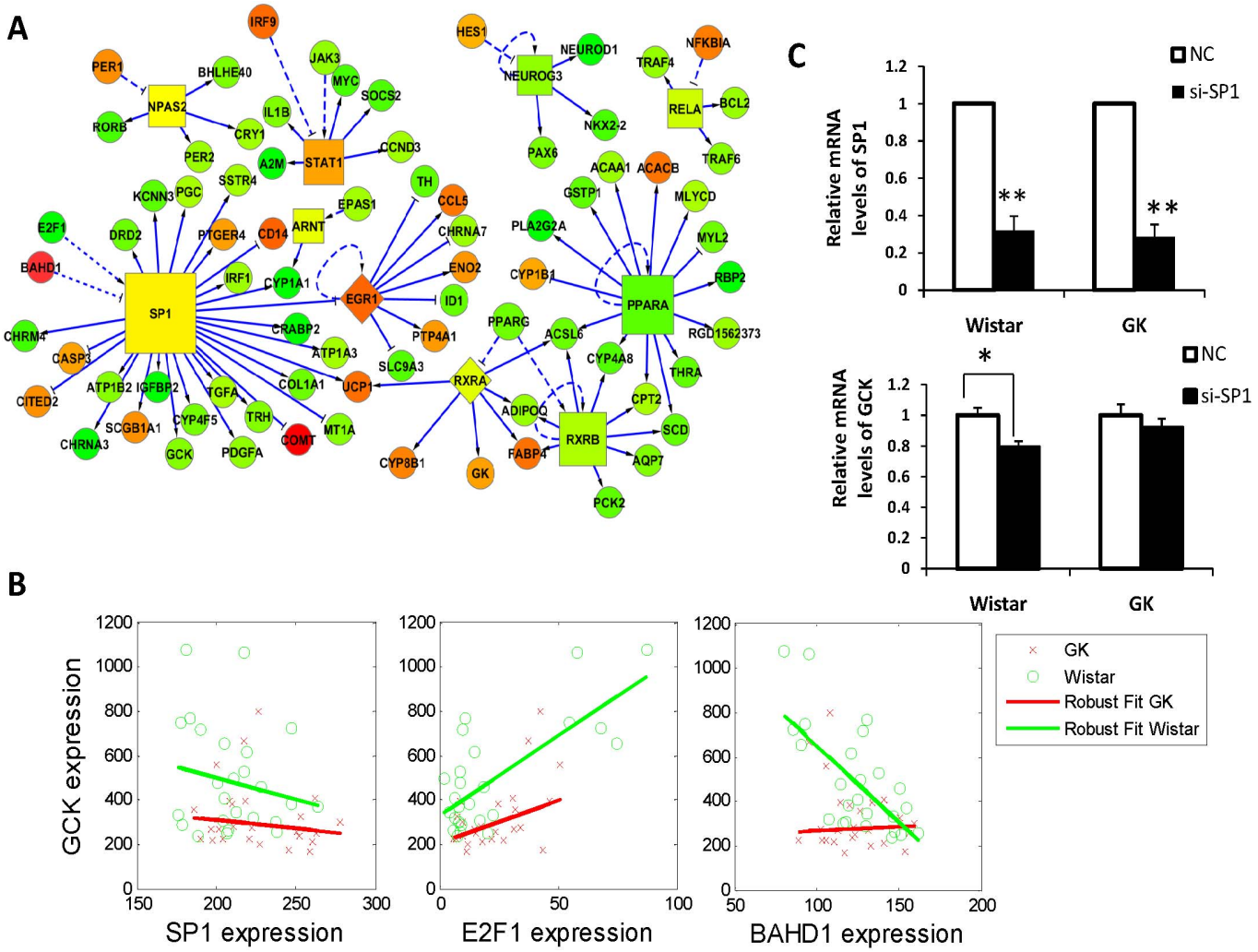

Figure 5 | (A) Network rewiring inferred by APG. Diamonds represent significantly active TFs, and squares represent significantly repressed TFs comparing GK to Wistar rats. The color of nodes ranging from green to red, represents mRNA expression levels from low to high compare GK with Wistar controls. All dash lines represent physical interaction between genes and TFs, that is the TF upstream-regulation, and all solid lines represent regulation between TFs and targets. (B) Robust fit showing the correlation between GCK expression and SP1, E2F1, BAHD1 in different rat models. (C) Knocking down SP1 decreased GCK mRNA expression in Wistar rats, but not in GK rats. The primary hepatocytes from Wistar and GK rats were transfected by negative control (NC) or SP1 siRNA for 24 hours. The GCK gene mRNA level was measured 24 hours after transfection. mRNA expression levels of GCK showed a remarkable decrease in Wistar hepatocytes after SP1 knocking down, while there was no significant change in GK hepatocytes after similar SP1 mRNA knocking down. Error bar indicates standard deviation for each group (n=3) in three separate experiments.

functions of gene network instead of gene list[34], and find that the predicted interactions are significantly enriched in immune system process, cellular response to chemical stimulus, and response to vitamin A and hormone stimulus, during all biological processes (Supplementary Table S1).

The APG network constructed above shows significantly differential TFs, cofactors, and target genes between GK and normal controls, within which SP1 is an obvious hub. A large number of growth and insulin responsive genes contain binding sites for the transcription factor SP1[24]. For instance, glucokinase (GCK), insulin-like growth factor binding protein 2 (IGFBP2), and PPAR-γ coactivator (PGC) are well known proteins and their dysfunctions are related to diabetes. From our previous published results and data, we also reported that SP1 regulates nuclear factors working in concert to keep normal glucose metabolism robustness[25]. The low activity of SP1 in GK rats may be associated with SP1 cofactors, that is, lower expression of E2F transcription factor 1 (E2F1) and higher expression of Bromo adjacent homology domain-containing protein 1 (BAHD1). SP1 and E2F1 act synergistically in activation of downstream gene transcription in transient transfection assays using Drosophila melanogaster SL2 cells[26]. BAHD1 is a novel heterochromatinization factor that leads to histone deacetylation and nucleosome compaction and reduces the ability of SP1 to promote transcription activities[27]. We observed that loss of SP1 activity caused significant increment of glucose production (Figure 4C). How SP1, as well as its potential cofactors E2F1 and BAHD1, affects glucose production was further investigated. We picked GCK, which is one target of SP1, discovered about 40 years ago independently in three laboratories[19,28,29]. Tissue survey indicates that the enzyme is liver specific, changing free glucose to glucose-6-phospharate for the synthesis of glycogen, thus serving a critical role in postprandial glucose clearance from the circulation[30,31]. In Supplementary Figure S9, GCK expression levels significantly decreased in diabetic GK rats, and GCK promoter clearly has SP1 binding region. As TF and target gene pairs, SP1 and GCK in expression levels (Figure 5B) were not significantly correlated in both normal Wistar and diabetic GK rats (p-values are 0.32 and 0.42 respectively based on a Student's t distribution for a transformation of the correlation). Instead, for the

first time, we found out that E2F1 and BAHD1, the potential cofactors of SP1, show strong correlations in expression levels with GCK in Wistar controls (p-values are 0.00003 for E2F1 and 0.0001 for BAHD1). Interestingly, all these correlations were dramatically weakened in GK rats compared with those in Wistar controls (p-values are 0.004 for E2F1 and 0.19 for BAHD1).

Those decreased correlations in GK rats are caused by low activity of SP1. We used RNA interference method to test regulation of GCK by SP1 in hepatocytes of GK and Wistar rats. If the activity of the TF is low in GK rat but high in Wistar rat, then the expression of the target genes should be little changed in the primary hepatocytes of GK rat but significant change in the primary hepatocytes of Wistar rat after knocking down by the specific siRNA to the TF. We detected the mRNA expression level of GCK in primary hepatocytes of Wistar rat and GK rat with or without SP1 knocking down by siRNA. As we expected, the real-time results showed that there is rarely change in GCK mRNA levels in hepatocytes of GK rat after SP1 siRNA knocking down. In contrast, GCK expression of mRNA level was significantly reduced in hepatocytes of Wistar rat after SP1 knocking down (**Figure 5C**).

Discussion

We developed a new mathematical method, named APG model, to quantitatively infer TF activities and network structures in an accurate manner from high throughput biological data. Specifically, in combination with gene expression levels, protein-protein interaction network, and transcription regulatory network, we integrated both TF upstream-regulation and downstream-regulation information to reveal transcriptional regulations of TFs at the network level. We theoretically prove that APG model has a unique solution for TF activity and is able to find unknown elements (i.e., unknown regulators and interactions) in the network structure even without enough prior knowledge, which significantly extends the previous methods in both theoretical and computational aspects. In addition, APG directly considers combinatorial interactions among TFs and cofactors, thereby enabling us to apply it to a wide class of biological data or solve real biological problems without treating each TF and cofactor in isolation. APG method with Gaussian graphical model, consists of three layers representing cofactors (cause layer), TF activity (hidden layer), and regulated target genes (effect layer) respectively, which can not only identify key factors in a biological process but also reveal hidden potential causal relations among biomolecules. The new approach was firstly validated by numerical experiments and then by biological experiments. Comparing with approaches only using downstream-regulation information, APG integrating multi-level information infers both TF activities and their network structure in a more accurate and robust manner. Moreover, applying our method in diabetic rats actually found hidden TFs with the changed activities during development of diseases, which cannot be identified by the traditional differential gene scheme due to no differential expression of those genes in microarray data. Actually, for many cases, those genes with high differential expression may be far from key regulators (or have even less relevance with the driving factors of the biological phenomenon) due to a cascade amplification effect of gene regulations from the key regulators, comparing to the genes with low differential expression. We further investigated the rewiring network of TFs and their regulatory targets in 4 week GK rats in comparison with Wistar controls, and for the first time revealed strong correlation of E2F1 and BAHD1 to GCK in Wistar rats. Our biological experiments validated SP1 as a hidden TF with changed regulatory activity, and the loss of SP1 activity was found to contribute to the increased glucose production during diabetes development in GK rats. Clearly, APG model provided opportunities to enhance our ability to use microarray data to elucidate transcriptional regulation in complex biological systems with combinatorial interactions of TFs and their cofactors.

Biological systems like individuals, cells, even mononuclear organisms are complex robust systems, in which TF combinations determine a specific phenomenon. Dr. Timothy screened for physical interactions of a large number of TFs and revealed highly interacted TFs conserved between mouse and human[32]. However, those methods only considering the significant changed mRNA or protein levels frequently miss important molecules which change their roles in the physiological or pathological process by rewiring TF interactions with their cofactors. On the other hand, theoretical models only considering TF downstream-regulation information may also miss main features of TF combinatorial interactions in biological systems. In contrast, APG model fully exploits TF upstream-regulation molecules, TFs, and TF downstream-regulation molecules, thereby providing us a powerful tool to quantify TF activities and infer the rewiring network of ''cofactors-TFs-regulatory targets'' behind a phenotype in an accurate manner by integrating high throughput data. Mapping the cofactors-TFs-regulatory targets interactions would significantly enhance our understanding of developmental processes and diseases. Applying APG to rodent high throughput data, followed by experimental confirmation, revealed important hidden roles of loss SP1 activity in development of diabetes in GK rats. Our data is consistent with clinical investigation, in which a dramatic reduction in Sp1 binding to GCK promoter sequence corresponds to GCK-MODY (maturity-onset diabetes of the young) cases[33]. Since the decreased GCK function causes mild fasting hyperglycemia, identifying mutations causing GCK hypofunction has important implications for treatment and prognosis; therefore, we propose to analyse E2F1 and BAHD1 in addition to SP1 for correct diagnosis of potential GCK-MODY. From our inferred network, SP1 also regulates important genes associated with metabolism, such as PGC and IGFBP2, et al. Currently, little is known about the direct role of SP1 in glucose metabolism of diabetes, partially due to similar mRNA levels of SP1 during diabetes progression. The vital function of SP1 in development and progression of diabetes has not gotten enough attention, and thus needs further investigation in the future.

Molecular experiments like chromatin immunoprecipitation (ChIP) and luciferase reporter assays provide us basic information about TF and target gene pairs. However, these in vitro experiments always treat each TF in isolation with unphysiological expression levels. Thus in a real physiological and pathological conditions, those regulation signals do not always exist. For example, the regulation of GCK by SP1 is lost in diabetic animals, as well as the strong correlations of BAHD1 to GCK. APG provides us a useful algorithm to calculate a specific regulatory signal in a complex system without treating each TF and cofactor in isolation. It has long been appreciated that combinatorial interactions among TFs and their cofactors change regulatory signals. Thus another significant contribution of the present work is to precisely calculate the regulatory signals on a global scale.

We found the relevance of loss of SP1 activity to diabetes in this study. E2F1 and BAHD1 dysfunctions are potential reasons reducing SP1 activity. However, it is also known that SP1 activities are involved in the complicated post-translational modifications, such as phosphorylations, proteolytic cleavage, glycosylation, and acetylation[24]. Thus we considered the mechanisms underlying loss of SP1 activity as our future topic. In addition, it is also our future work to study TF activities for biological or medical problems by considering dynamical features[35,36] of disease progression.

Methods

We first collected the gene expression data of TFs, potential cofactors and their target genes from GK and Wistar rats in Gene Expression Omnibus (GEO) database (GSE 13271). There were totally 50 samples with different phenotype (diabetes and normal control) in five different time points. We further selected targeted genes in the data set, by setting the threshold of the fold change of gene expressions between GK and Wistar rats (fold change >1.5 or $<1/1.5$). Then, the prior knowledge including protein-protein interaction networks, and TF-gene regulatory networks to construct connectivity pattern (potential interaction network) was prepared from KEGG[15],

TRED database[16] and homologous from human and mouse. Then we obtained 905 potential regulations between TFs and their targets, and among them, 468 pairs are further known about their regulatory directions (activation or repression of target expression). We also obtained 326 potential protein-TF interactions, and among them, 263 pairs are known about regulatory directions (activation or repression of TF activity). Finally, 113 TFs, 121 TF-interacting proteins, and 335 targets with both network and expression information were selected for further research.

For the purpose to experimentally validate our results, male Wistar and GK rats at 4 weeks of age were purchased from SLRC Laboratory center. Animals were housed in cages at a constant temperature with 12 hours light/dark cycles. Animals were given free access to water and food. All experiments were approved by Institutional Animal Care and Use Committee at shanghai Institutes of Biological Science, China Academy of Sciences. Livers from fed Wistar or GK rats were washed, and further digested with 0.04% collagenase (Worthington, LS004196). Cells with viability higher than 85% as assessed by the trypan blue exclusion test were used for experiments. The primary hepatocytes were cultured in DMEM medium with low glucose (1 g/L) (Invitrogen) supplemented with 10% fetal bovine serum, 100 U/ml penicillin and 100 ug/ml streptomycin. We performed siRNA knockdown by transient transfection. Duplex siRNA oligonucleotides were synthesized by GenePharma. Ltd (Shanghai, China). The RFect transfection reagent was purchased from Bio-Tran Biotechnologies (Shanghai, China). At 24 hours after transfection of siRNA, the hepatocytes were gently washed with PBS three times, followed by incubation with glucose production buffer (DMEM with no glucose but supplemented with 20 mM sodium lactate and 2 mM sodium pyruvate). Supernatant was collected for measuring glucose concentration by fluorometric assay (Invitrogen). The results were further normalized with total protein in whole-cell lysates. For detailed experimental methods, refer to the Supplementary Information.

1. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
2. Domon, B. & Aebersold, R. Mass spectrometry and protein analysis. *Science* **312**, 212–217 (2006).
3. Joung, J. K., Ramm, E. I. & Pabo, C. O. A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 7382–7387 (2000).
4. Mann, M., Ong, S. E., Gronborg, M., Steen, H., Jensen, O. N. & Pandey, A. Analysis of protein phosphorylation using mass spectrometry: deciphering the phosphoproteome. *Trends in biotechnology*, **20**, 261–268 (2002).
5. Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends in genetics: TIG* **24**, 133–141 (2008).
6. Liang, P. & Pardee, A. B. Analysing differential gene expression in cancer. *Nature reviews Cancer* **3**, 869–876 (2003).
7. Barabasi, A. L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. Nature reviews. *Genetics* **5**, 101–113 (2004).
8. Friedman, A. & Perrimon, N. Genetic screening for signal transduction in the era of network biology. *Cell* **128**, 225–231 (2007).
9. Kitano, H. Systems biology: a brief overview. *Science* **295**, 1662–1664 (2002).
10. Liao, J. C., Boscolo, R., Yang, Y. L., Tran, L. M., Sabatti, C. & Roychowdhury, V. P. Network component analysis: reconstruction of regulatory signals in biological systems. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 15522–15527 (2003).
11. Carro, M. S. *et al*. The transcriptional network for mesenchymal transformation of brain tumours. *Nature* **463**, 318–325 (2010).
12. Almon, R. R., DuBois, D. C., Lai, W., Xue, B., Nie, J. & Jusko, W. J. Gene expression analysis of hepatic roles in cause and development of diabetes in Goto-Kakizaki rats. *The Journal of endocrinology* **200**, 331–346 (2009).
13. Kitahara, A., Toyota, T., Kakizaki, M. & Goto, Y. Activities of hepatic enzymes in spontaneous diabetes rats produced by selective breeding of normal Wistar rats. *The Tohoku journal of experimental medicine* **126**, 7–11 (1978).
14. Surtees, P. G., Wainwright, N. W. & Gilks, W. R. Diagnostic complexity and depression: time to allow for uncertainty. *Psychological medicine* **26**, 1105–1110 (1996).
15. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research* **40**, D109–114 (2012).
16. Zhao, F., Xuan, Z., Liu, L. & Zhang, M. Q. TRED: a Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies. *Nucleic acids research* **33**, D103–107 (2005).
17. Kadonaga, J. T. & Tjian, R. Affinity purification of sequence-specific DNA binding proteins. *Proceedings of the National Academy of Sciences of the United States of America* **83**, 5889–5893 (1986).
18. Kavurma, M. M. & Khachigian, L. M. Sp1 inhibits proliferation and induces apoptosis in vascular smooth muscle cells by repressing p21WAF1/Cip1 transcription and cyclin D1-Cdk4-p21WAF1/Cip1 complex formation. *The Journal of biological chemistry* **278**, 32537–32543 (2003).
19. Sols, A., Salas, M. & Vinuela, E. Induced biosynthesis of liver glucokinase. *Advances in enzyme regulation* **2**, 177–188 (1964).
20. Nielsen, J. H. *et al*. Regulation of beta-cell mass by hormones and growth factors. *Diabetes* **50 Suppl 1**, S25–29 (2001).
21. Davoodi-Semiromi, A. *et al*. A Mutant Stat5b with Weaker DNA Binding Affinity Defines a Key Defective Pathway in Nonobese Diabetic Mice. *The Journal of biological chemistry* **279**, 11553–11561 (2004).
22. Brown, N. S. *et al*. Thyroid hormone resistance and increased metabolic rate in the RXR-gamma-deficient mouse. *The Journal of clinical investigation* **106**, 73–79 (2000).
23. Movassat, J., Saulnier, C., Serradas, P. & Portha, B. Impaired development of pancreatic beta-cell mass is a primary event during the progression to diabetes in the GK rat. *Diabetologia* **40**, 916–925 (1997).
24. Samson, S. L. & Wong, N. C. Role of Sp1 in insulin regulation of gene expression. *Journal of molecular endocrinology* **29**, 265–279 (2002).
25. Zhou, H. *et al*. Network screening of Goto-Kakizaki rat liver microarray data during diabetic progression. *BMC systems biology* **5 Suppl 1**, S16 (2011).
26. Lin, S. Y. *et al*. Cell cycle-regulated association of E2F1 and Sp1 is related to their functional interaction. *Molecular and cellular biology* **16**, 1668–1675 (1996).
27. Bierne, H. *et al*. Human BAHD1 promotes heterochromatic gene silencing. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 13826–13831 (2009).
28. Matschinsky, F. M. & Ellerman, J. E. Metabolism of glucose in the islets of Langerhans. *The Journal of biological chemistry* **243**, 2730–2736 (1968).
29. Grimsby, J. *et al*. Allosteric activators of glucokinase: potential role in diabetes therapy. *Science* **301**, 370–373 (2003).
30. Vandercammen, A. & Van Schaftingen, E. The mechanism by which rat liver glucokinase is inhibited by the regulatory protein. *European journal of biochemistry/FEBS* **191**, 483–489 (1990).
31. Magnuson, M. A., Andreone, T. L., Printz, R. L., Koch, S. & Granner, D. K. Rat glucokinase gene: structure and regulation by insulin. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 4838–4842 (1989).
32. Ravasi, T. *et al*. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* **140**, 744–752 (2010).
33. Gasperikova, D. *et al*. Identification of a novel beta-cell glucokinase (GCK) promoter mutation (−71G>C) that modulates GCK gene expression through loss of allele-specific Sp1 binding causing mild fasting hyperglycemia in humans. *Diabetes* **58**, 1929–1935 (2009).
34. Wang, J. *et al*. NOA: a novel network ontology analysis method. *Nucleic Acids Res* **39**, e87 (2011).
35. Chen, L. *et al*. Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. *Scientific Reports* **2**, 342 (2012).
36. He, D. *et al*. Coexpression network analysis in chronic hepatitis B and C hepatic lesion reveals distinct patterns of disease progression to hepatocellular carcinoma. *Journal of Molecular Cell Biology* **4** (3), 140–52 (2012).

Acknowledgement

We thank Katsuhisa Horimoto and Qianfei Wang for critical reading and suggestions. This work was supported by grants from Major State Basic Research Development Program of China (973 Program) under No.2011CB504003 (H.Z.); National Natural Science Foundation of China under No. 61134013 (H.Z., L.C.), No. 81070657 (H.Z.), Nos.61072149, 91029301 (L.C.), and No.11131009 (J.W., X.Z.); NN-CAS Research Foundation under No. NNCAS-2009-1 (H.Z.); Shanghai Pujiang Program (L.C.); Chief Scientist Program of Shanghai Institutes for Biological Sciences, CAS under No. 2009CSP002 (L.C.); the FIRST program from JSPS initiated by CSTP (L.C.).

Author contributions

J.W. and L.C. designed research; J.W. performed the 'dry-lab' research; Y.S. and H.Z. designed and performed the 'wet-lab' research; J.W. and Y.S. analyzed data; J.W., L.C. and H.Z. wrote the paper; and all authors revised the paper.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Wang, J. *et al*. APG: an Active Protein-Gene Network Model to Quantify Regulatory Signals in Complex Biological Systems. *Sci. Rep.* **3**, 1097; DOI:10.1038/srep01097 (2013).